Systematic Gene Search in the Incyte LifeSeq Database

Normal tissue
~50,000 individual ESTs

Tumor tissue
~50,000 individual ESTs

Priority list
High
- Prostate
- Breast
- Ovary
- Bladder
- Uterus

Low

Iterative assembling with increasing mismatch

~8,000 contigs
+
~25,000 individual sequences

~8,000 contigs
+
~25,000 individual sequences

Comparison of databases

normal tissue-specific (expected: 100-500)

nonspecifically expressed genes

tumor tissue-specific (expected: 100-500)

Genes of Interest

Figure 1

Systematische Gen-Suche in der Incyte LifeSeq Datenbank

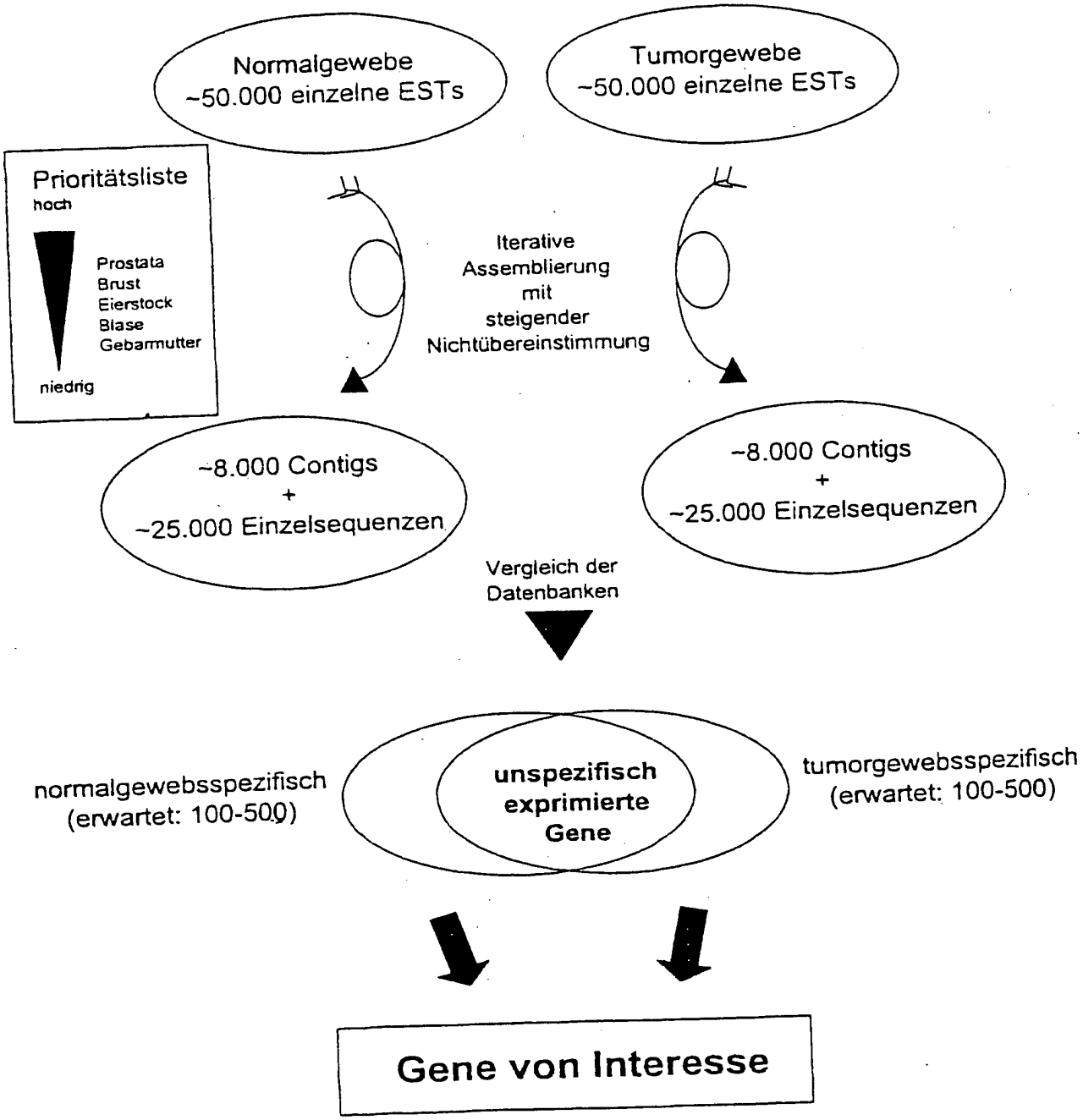


Fig. 1

Principle of EST Assembly

~50,000 ESTs per tissue

Assembly at 0% mismatch with GAP4 (Staden)

Contigs

Individual Sequences

Contigs increasing in number and length

Iterative assembly with increasing mismatch (1%, 2%, 4%)

5000-6000 contigs

~25,000 other individual sequences

~30,000 consensus-sequences per tissue

Figure 2a

Prinzip der EST-Assemblierung

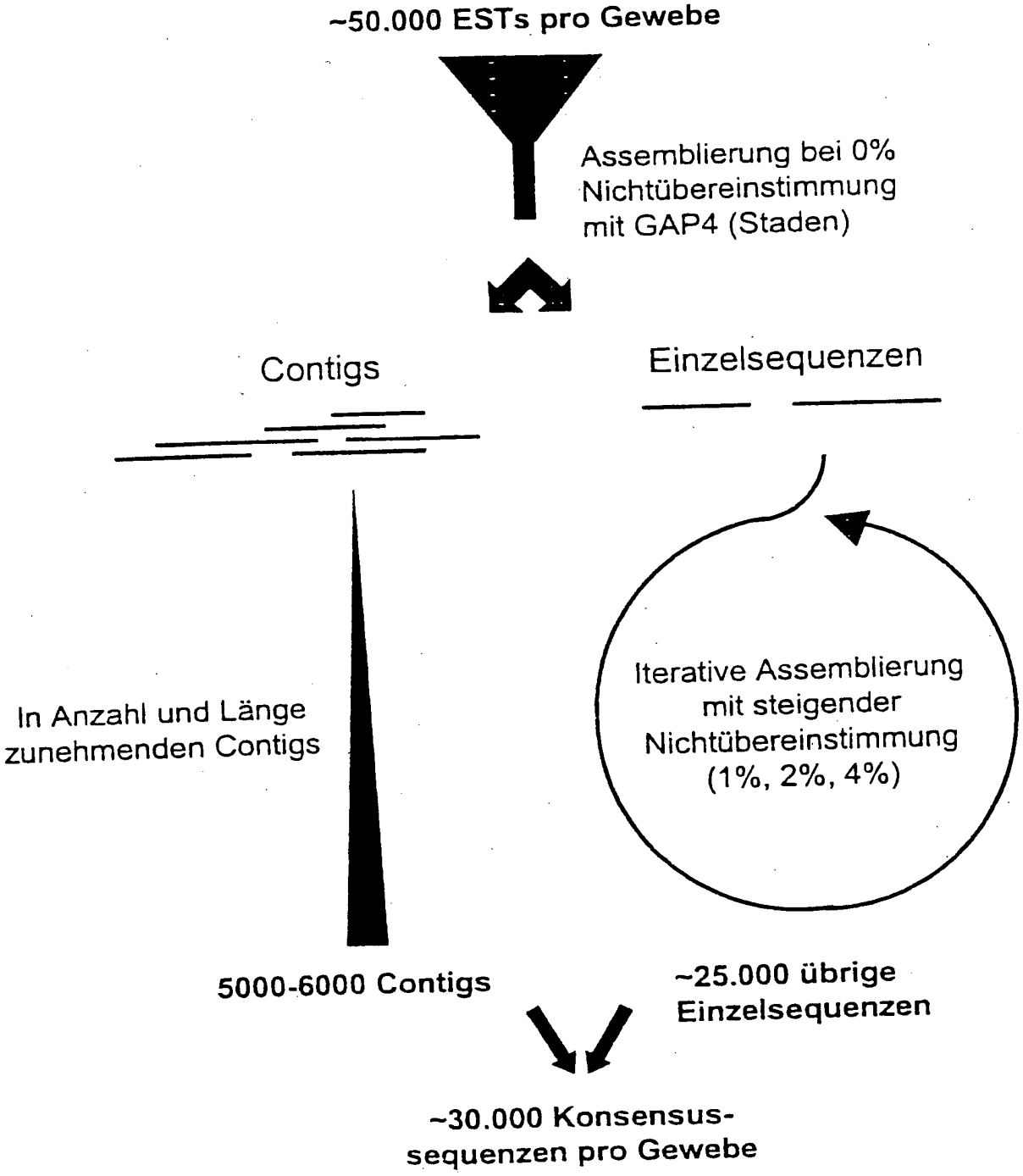


Fig. 2a

~50,000 ESTs of a tissue (e.g.: uterus tumor)

GAP4 Assembly 1st Round:
minimum initial match: 20
maximum number of inserted blanks per sequence: 8
maximum percent mismatch: 0

GAP4 Database 1
Contigs 1
Individual sequences 1

unassembled
ESTs

GAP4 Assembly 2nd Round:
minimum initial match: 20
maximum number of inserted blanks per sequence: 8
maximum percent mismatch: 1

GAP4 Database 2
Contigs 2
Individual sequences 2

unassembled
ESTs

GAP4 Assembly 3rd Round:
minimum initial match: 20
maximum number of inserted blanks per sequence: 8
maximum percent mismatch: 2

GAP4 Database 3:
Contigs 3
Individual sequences 3

unassembled
ESTs

Figure 2b1

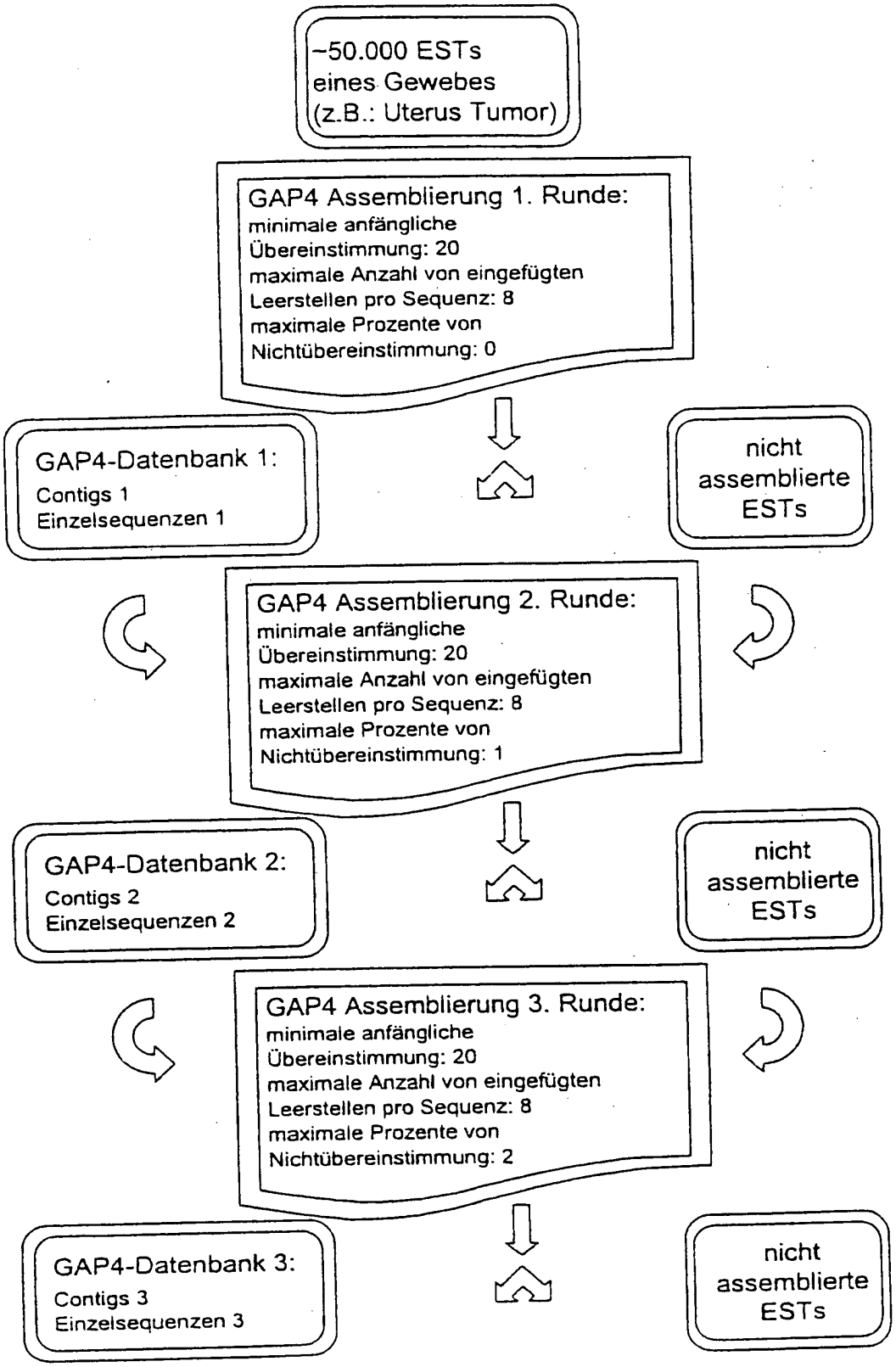

~50.000 ESTs
eines Gewebes
(z.B.: Uterus Tumor)
GAP4 Assemblierung 1. Runde:
minimale anfängliche
Übereinstimmung: 20
maximale Anzahl von eingefügten
Leerstellen pro Sequenz: 8
maximale Prozente von
Nichtübereinstimmung: 0
GAP4-Datenbank 1:
Contigs 1
Einzelsequenzen 1
nicht
assemblierte
ESTs
GAP4 Assemblierung 2. Runde:
minimale anfängliche
Übereinstimmung: 20
maximale Anzahl von eingefügten
Leerstellen pro Sequenz: 8
maximale Prozente von
Nichtübereinstimmung: 1
GAP4-Datenbank 2:
Contigs 2
Einzelsequenzen 2
nicht
assemblierte
ESTs
GAP4 Assemblierung 3. Runde:
minimale anfängliche
Übereinstimmung: 20
maximale Anzahl von eingefügten
Leerstellen pro Sequenz: 8
maximale Prozente von
Nichtübereinstimmung: 2
GAP4-Datenbank 3:
Contigs 3
Einzelsequenzen 3
nicht
assemblierte
ESTs

Fig. 2b1

GAP4 Database 3:
Contigs 3 Individual Sequences 3

unassembled ESTs

Consensus 3

GAP4 Assembly 4th Round:
minimum initial match: 20
maximum number of inserted blanks per sequence: 8
maximum percent mismatch: 2

GAP4 Database 4:
Contigs 4 Individual Sequences 4

unassembled ESTs

Consensus 4

GAP4 Assembly 5th Round:
minimum initial match: 20
maximum number of inserted blanks per sequence: 8
maximum percent mismatch: 4

GAP4 Database 5:
Contigs 5 Individual Sequences 5

unassembled ESTs 5

Consensus 5

Individual Sequences 5

Figure 2b2

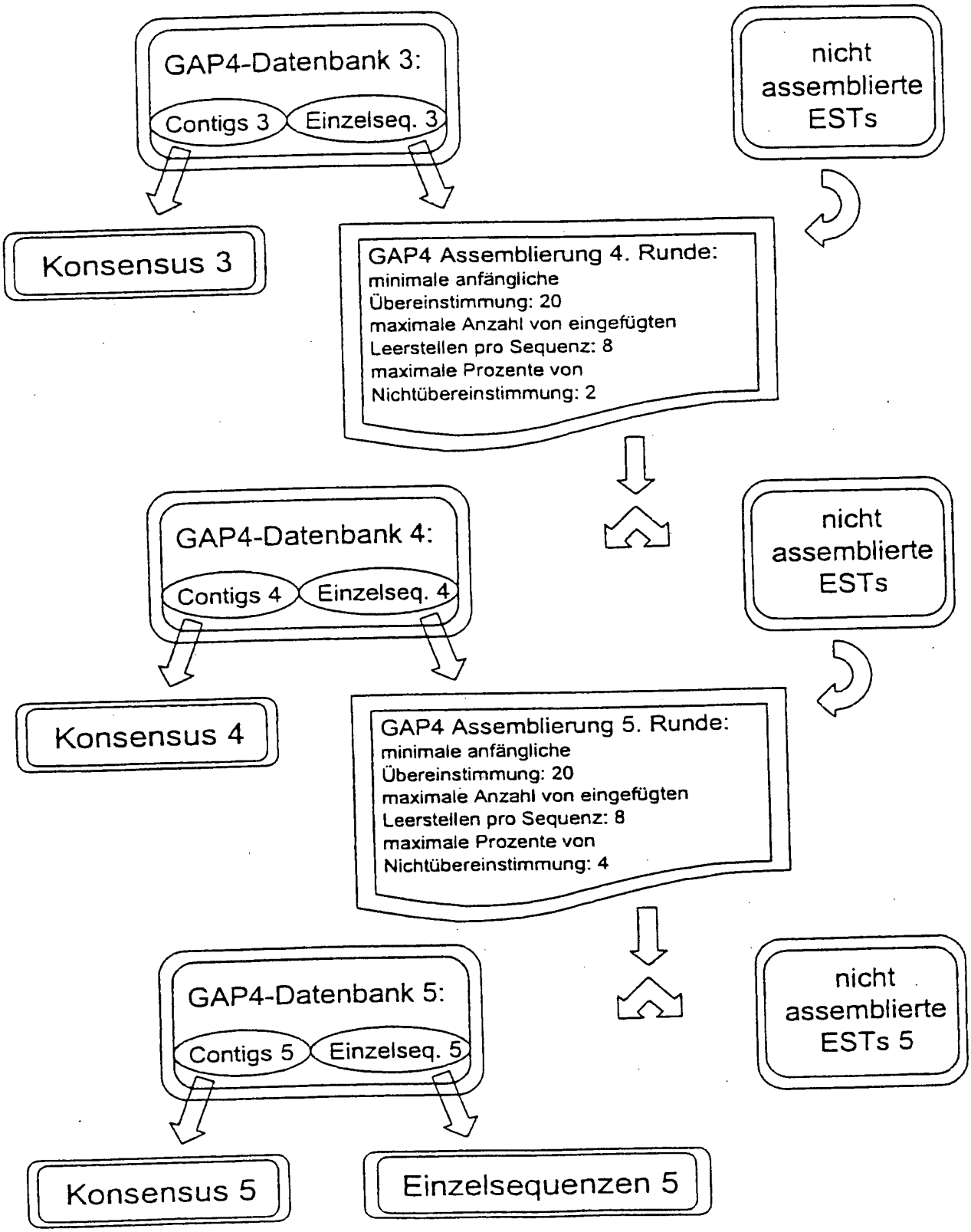

GAP4-Datenbank 3:
Contigs 3
Einzelseq. 3
nicht assemblierte ESTs
Konsensus 3
GAP4 Assemblierung 4. Runde:
minimale anfängliche Übereinstimmung: 20
maximale Anzahl von eingefügten Leerstellen pro Sequenz: 8
maximale Prozente von Nichtübereinstimmung: 2
GAP4-Datenbank 4:
Contigs 4
Einzelseq. 4
nicht assemblierte ESTs
Konsensus 4
GAP4 Assemblierung 5. Runde:
minimale anfängliche Übereinstimmung: 20
maximale Anzahl von eingefügten Leerstellen pro Sequenz: 8
maximale Prozente von Nichtübereinstimmung: 4
GAP4-Datenbank 5:
Contigs 5
Einzelseq. 5
nicht assemblierte ESTs 5
Konsensus 5
Einzelsequenzen 5

Fig. 2b2

Consensus 3 Individual Sequences 5

Consensus 4 unassembled ESTs 5

Consensus 5

GAP4 Assembly 6th Round:
minimum initial match: 20
maximum number of inserted blanks per sequence: 8
maximum percent mismatch: 4

Assembled database
of a specific tissue
(e.g.: uterus tumor)

Figure 2b3

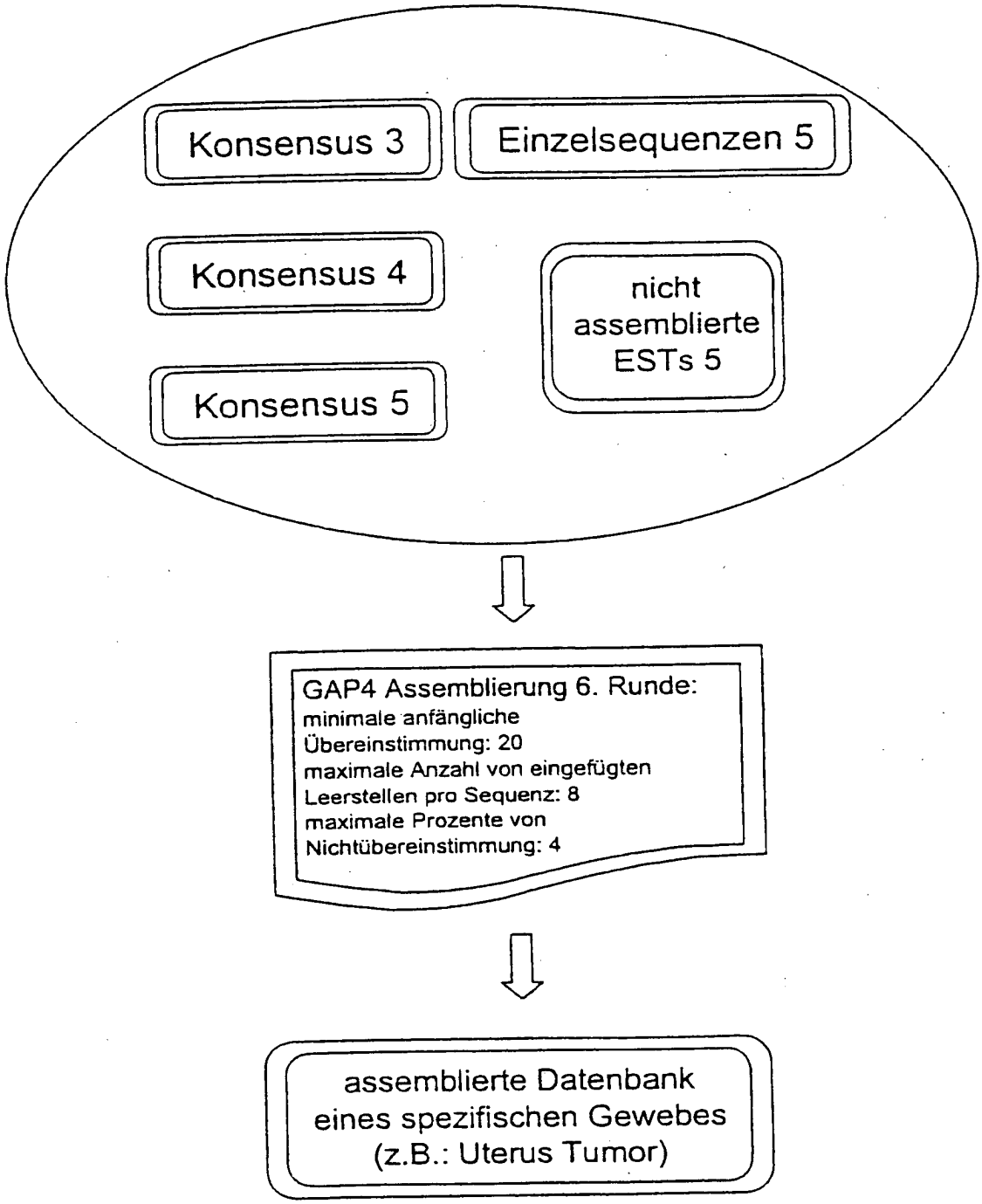

Konsensus 3
Einzelsequenzen 5
Konsensus 4
nicht assemblierte ESTs 5
Konsensus 5
GAP4 Assemblierung 6. Runde:
minimale anfängliche Übereinstimmung: 20
maximale Anzahl von eingefügten Leerstellen pro Sequenz: 8
maximale Prozente von Nichtübereinstimmung: 4
assemblierte Datenbank eines spezifischen Gewebes (z.B.: Uterus Tumor)

Fig. 2b3

Assembled database
of a specific tissue
(e.g.: uterus tumor)

Consensus 6

Read-in as individual sequences

Database
of a specific tissue
(e.g.: uterus tumor)

Database of a second
specific tissue
(e.g.: normal uterus)

GAP4 Assembly
minimum initial match: 20
maximum number of inserted blanks
per sequence: 8
maximum percent mismatch: 4

Tumor tissue-
specific ESTs

Non-tissue-
specific ESTs

Normal tissue-
specific ESTs

Fig. 2b4

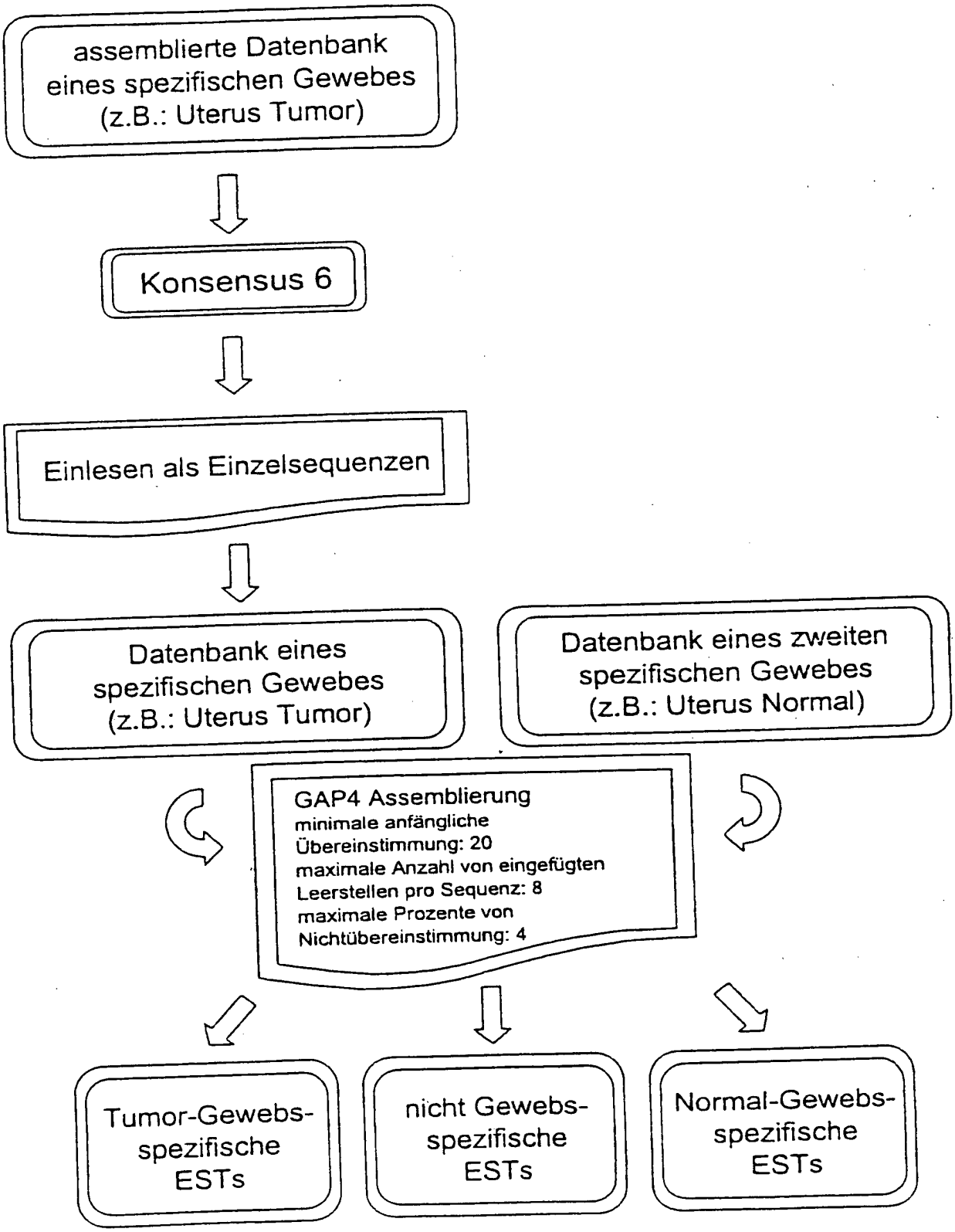

assemblierte Datenbank eines spezifischen Gewebes (z.B.: Uterus Tumor)
Konsensus 6
Einlesen als Einzelsequenzen
Datenbank eines spezifischen Gewebes (z.B.: Uterus Tumor)
Datenbank eines zweiten spezifischen Gewebes (z.B.: Uterus Normal)
GAP4 Assemblierung
minimale anfängliche Übereinstimmung: 20
maximale Anzahl von eingefügten Leerstellen pro Sequenz: 8
maximale Prozente von Nichtübereinstimmung: 4
Tumor-Gewebs-spezifische ESTs
nicht Gewebs-spezifische ESTs
Normal-Gewebs-spezifische ESTs

Fig. 2b4

In silico subtraction of gene expression in various tissues

~30,000 consensus sequences
normal tissue

~30,000 consensus sequences
tumor tissue

Assembly at 4% mismatch

Normal tissue
Specific genes

Cancer tissue
Specific genes

Genes expressed in both tissues

Figure 3

CORRECTED PAGE (RULE 91)
ISA/EP

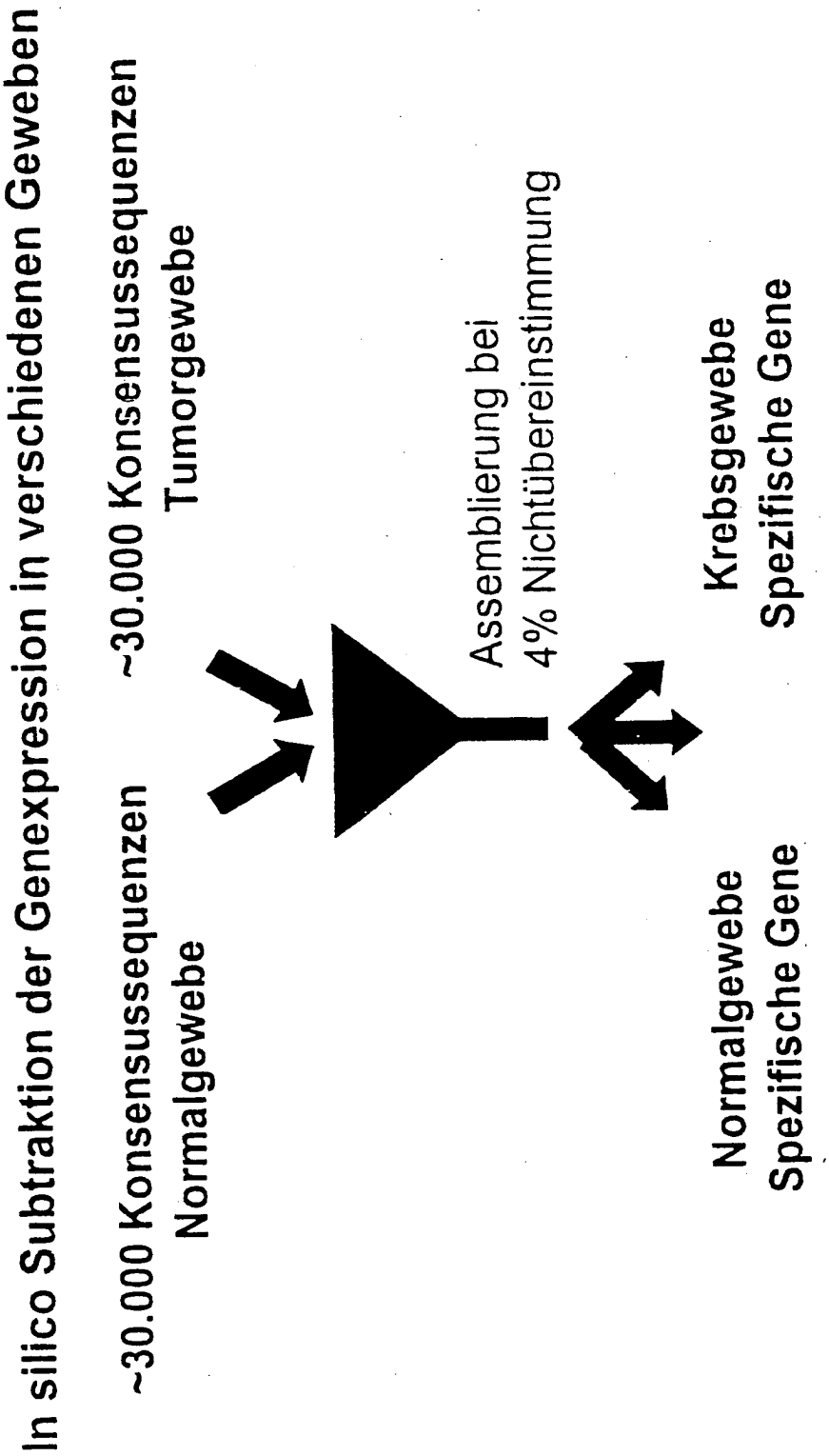

In silico Subtraktion der Genexpression in verschiedenen Geweben
~30.000 Konsensussequenzen
Normalgewebe
~30.000 Konsensussequenzen
Tumorgewebe
Assemblierung bei
4% Nichtübereinstimmung
Normalgewebe
Spezifische Gene
Krebsgewebe
Spezifische Gene
In beiden Geweben
exprimierte Gene

Fig. 3

Genes of interest

Determination of tissue-specific expression via electronic Northern (INCYTE LifeSeq and public EST databases)

Candidate genes for tumor suppressors or tumor activators

Figure 4a

CORRECTED PAGE (RULE 91)
ISA/EP

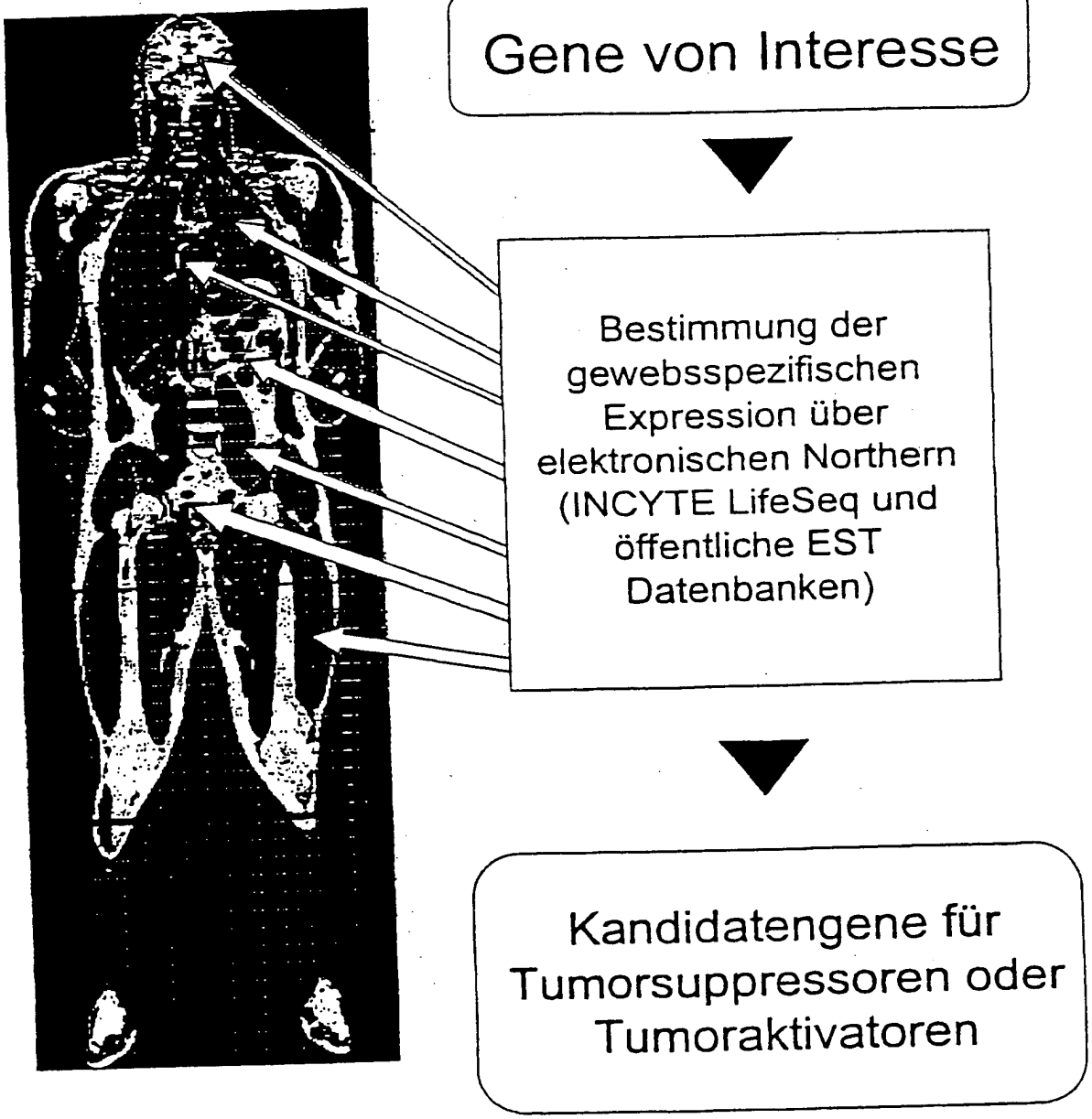

Gene von Interesse
Bestimmung der gewebsspezifischen Expression über elektronischen Northern (INCYTE LifeSeq und öffentliche EST Datenbanken)
Kandidatengene für Tumorsuppressoren oder Tumoraktivatoren

Fig. 4a

Partial cDNA sequence
e.g., EST or contig
S

...GCCTCAAGTTATC...

WHILE $C_i > C_{i-1}$

Electronic Northern Blot

Fisher's Exact Test IF H_0 EXIT

Automatic Lengthening

Consensus sequence C

...ATGTCCTAGCCTCAAGTTATCAGATGCAA...

Figure 4b

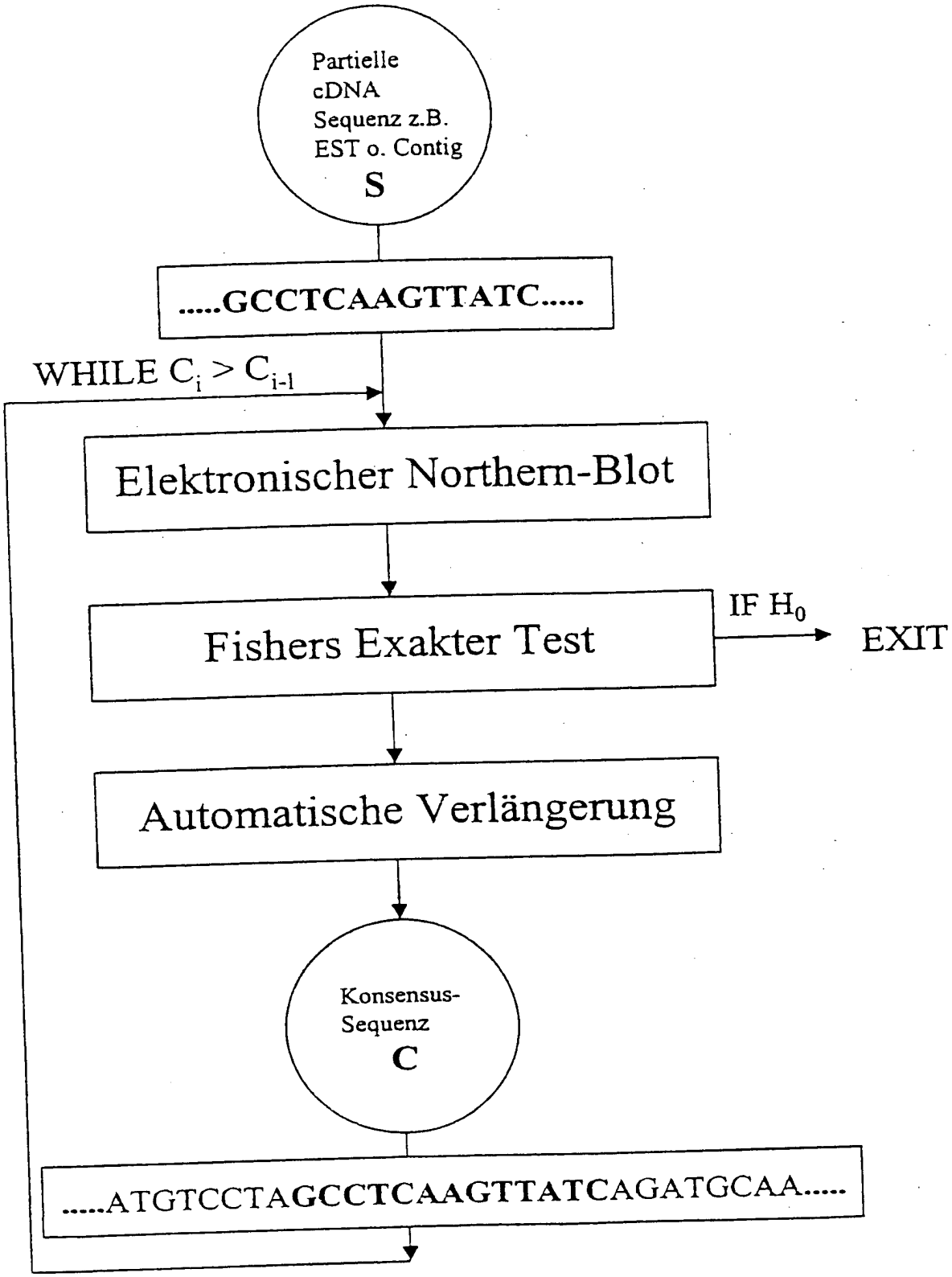

Partielle
cDNA
Sequenz z.B.
EST o. Contig
S
.....GCCTCAAGTTATC.....
WHILE $C_i > C_{i-1}$
Elektronischer Northern-Blot
Fishers Exakter Test
IF H_0
EXIT
Automatische Verlängerung
Konsensus-
Sequenz
C
.....ATGTCCTAGCCTCAAGTTATCAGATGCAA.....

Fig. 4b

Isolation of genomic BAC and PAC clones

Chromosomal clone localization via FISH

Hybridization signal

Sequencing of clones that are located in regions that have chromosomal deletions in prostate and breast cancer leads to identification of candidate genes

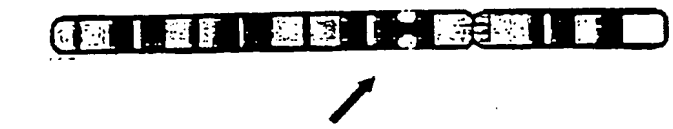
Exon Intron

Confirmation of candidate genes by screening of mutations and/or deletions in cancer tissues

Figure 5

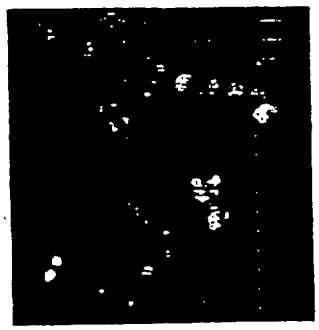
CORRECTED PAGE (RULE 91)
ISA/EP

Isolieren von genomischen BAC und PAC Klonen

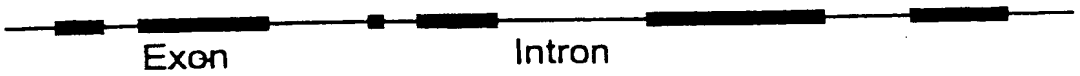


Chromosomale Klon-Lokalisation über FISH

Hybridisierungssignal



Sequenzierung von Klonen, die in Regionen lokalisiert sind, die chromosomale Deletionen in Prostata- und Brustkrebs aufweisen, führt zur Identifizierung von Kandidatengenen



Bestätigung der Kandidatengene durch Screening von Mutationen und/oder Deletionen in Krebsgeweben

Fig. 5